COMPUTATIONAL STUDIES OF PROTEIN FOLDING

The authors describe the state of the art in the field of protein structure prediction. They also introduce Prospector, a newly developed, iterative threading algorithm for protein structure prediction that can also be applied to ab initio protein folding, and discuss the promising results of its large-scale application.

Proteins are the workhorses of life. These polymers, comprised of 20 naturally occurring amino acids, fold to a unique, biologically active conformation called the *native state*. Various genome-sequencing projects now list the parts of such protein sequences in a given organism, but unfortunately, this list is of little utility; the real need is to identify the functions of all these proteins, which range from molecular to physiological to phenotypical. For between 40 to 60 percent of the protein-coding regions (or open reading frames), sequence-based methods that exploit evolutionary information can provide insight into some aspect of biological function. Such alignment methods define the standard against which we must measure all alternative approaches,[1,2] but such approaches increasingly fail as the protein families become more distant.[3] The remaining unassigned open reading frames represent an important challenge, and structure-based approaches to function prediction can play a significant role,[3,4] especially in target selection for genomics projects.[5] The ultimate goal for most such projects is to experimentally determine the structure of all possible protein folds so that any newly found sequence is within modeling distance of an already solved structure. In this article, we examine the status of contemporary protein structure prediction approaches.

There are three classes of approaches to protein structure prediction: homology modeling, threading, and ab initio folding. In homology modeling,[6] the query sequence and the already solved template structure are clearly evolutionarily related. The key challenge is to generate the best alignment on the template backbone, rebuild the protein's side chains, and fill in the alignment's gaps, typically in the loops between secondary structural elements. For threading, we attempt to find the closest matching structure in a library of already solved structures.[7] The structures can be analogous—two proteins need not be evolutionarily related—but they adopt similar structures by convergent evolution. Both threading and homology modeling have the disadvantage that a solved example of a related structure must already exist. With ab initio folding, we attempt to fold a protein from a random conformation.[8] It has the advantage that we don't need a previous example of the fold, but it

1521-9615/01/$10.00 © 2001 IEEE

JEFFREY SKOLNICK
Danforth Plant Science Center, Missouri
ANDRZEJ KOLINSKI
Danforth Plant Science Center and the University of Warsaw, Poland

is limited to relatively small proteins and generally produces low- to moderate-resolution predicted structures.

Ab initio folding has a particular impact on two problems that we must simultaneously solve for ab initio protein-structure prediction to be truly successful. We must first develop an efficient conformational search scheme that addresses the multiple minima problem (if each residue has three degrees of freedom, then a 100-residue protein has on the order of 10^{50} possible conformations), and then we must apply it to an energy function that has the global minimum in the protein's native conformation. Both parts are equally challenging, because a protein's energy landscape has many hills and valleys, and developing an energy function that identifies the native state as the global minimum among similar, but incorrect, protein-like states is nontrivial. These problems can be partly addressed by exploiting information from threading such as predicted contacts between side chains. In this spirit, we have developed a unified approach to protein structure prediction that uses information from our new threading algorithm Prospector as restraints in an ab initio folding algorithm.

A historical perspective

A typical protein folds from the unfolded, random conformation state to the native state on the order of milliseconds to minutes. At full atomic detail, we would have to simulate both the protein and the water in which it is dissolved. Using contemporary computers, it is impossible to fold a protein by brute force. Classical molecular dynamics simulations of a protein surrounded by an appropriate number of water molecules typically access times on the order of tens to hundreds of nanoseconds, which is at least three orders of magnitude less than the fastest protein folding times. To simulate the requisite folding time scales, we typically reduce the number of the protein's degrees of freedom and treat the solvent implicitly by a potential of mean force (such as a Generalized Born (which treats the electrostatics), accessible surface approach.[9]

First steps

The first reduced protein folding models appeared 25 years ago. In their pioneering work, Michael Levitt and Arieh Warshel proposed a model that assumed two centers of interaction per residue, one on the backbone alpha carbon and the other at the side group mass's center.[10] Each amino acid had a single degree of freedom involving its rotation around the Cα-Cα virtual bond. A knowledge-based potential controlled the short-range interactions, while the interactions between the side groups were of the Lennard-Jones type. They handled sampling with classical molecular dynamics. Simulations of bovine pancreatic trypsin inhibitor sometimes produced structures resembling the native fold, with the best structures having a root-mean-square deviation from the native in the range of 6.5 Å.

Later researchers studied similar models, with comparable results.[11] Some have developed continuous-space models with more structural details. Sun examined models that had an all-atom representation of the main chain and single, united atom-side groups.[12] Knowledge-based statistical potentials described the interactions between the side groups, and a genetic algorithm (GA) searched conformational space. For small peptides (such as mellitin, pancreatic polypeptide inhibitor, and apamin), he predicted structures whose accuracy ranged from a root-mean-square deviation of 1.66 Å to 4.5 Å from native, depending on size.Pedersen and John Moult assumed an all-heavy atom protein representation and used knowledge-based potentials to describe intraprotein interactions.[13] A combination of Monte Carlo (MC) and GAs search the conformational space. MC produces a set of structures for the GA starting population, with crossover points occurring in the largest flexibility regions detected in the MC runs. Their method successfully predicted low- to moderate-resolution protein fragments and the approximate folds of small proteins, but it's limited to small proteins.

There are three classes of approaches to protein structure prediction.

Lattices to simplify the conformational search

Although continuous-space, reduced models contain fewer degrees of freedom than detailed atomic models, effectively sampling the conformational space for larger proteins is extremely difficult. To further reduce the number of degrees of freedom, researchers have proposed discrete or lattice models. Early studies of lattice proteins did not focus on protein structure prediction but rather on understanding the fundamentals of the thermodynamics and kinetics of protein folding.[14–21]

The first attempt to predict a protein's native

structure in an ab initio fashion using a lattice representation of a protein came from Dashevskii.[22] He used a diamond lattice chain to approximate the polypeptide conformations and a chain growth algorithm to sample conformational space. A simple force field generated and identified compact structures resembling native folds of small polypeptides. Somewhat later, David Covell investigated a simple cubic lattice model of real proteins.[23] His model's force field consisted entirely of long-range interactions that included a pairwise, knowledge-based potential, a surface term, and a potential that corrected the model chain's local packing. The quality of the crude folds this method generated was comparable to those obtained from early continuous models. Covell later studied five small globular proteins by the enumeration of all possible compact conformations on a body-centered cubic lattice chain. He and his colleagues could always find the closest-to-native conformation within the top 2 percent of the lowest energy structures, as assessed by a knowledge-based interaction scheme.

Such models generated correct low- to moderate-resolution structures.

Hinds and Michael Levitt developed a lattice model of proteins where a single diamond lattice vertex represented several residues of a real protein.[24] They used an elaborate statistical potential to mimic the mean interactions between such defined protein segments and did an exhaustive search of a compact space. They then obtained the actual identity of the residues from a dynamic programming procedure. Often, they found correct low-resolution structures among the compact structures.

Over the years, we (the authors) have developed a series of high-coordination lattice models of globular proteins.[17,28,25,26] We used lattices of various resolution to mimic the Cα-trace of real proteins, ranging from 3D "chess-knight" type lattices to a high coordination lattice with 90 lattice vectors to represent possible subsequent locations of Cα-Cα virtual bonds. The models had additional interaction centers to represent the side groups, described by a single-sphere, multiple-rotamer representation.[26,27] The force field contained terms mimicking short-range interactions that described local conformational preferences for helices and beta strands; explicitly cooperative hydrogen bonds; and one body, pairwise, and multibody long-range interactions, with an implicit averaged effect of the water molecules. For several small globular proteins and simple multimeric coils, such models generated correct low- to moderate-resolution (high-resolution in the case of leucine zippers) structures obtained from simulated annealing simulations.[26,28]

CASP

To assess the current status of protein structure prediction, John Moult proposed the CASP (Critical Assessment of Techniques for Protein Structure Prediction) community-wide protein structure prediction experiment. The idea is that experimentalists who are about to determine protein structures make the sequences of the proteins available and then the protein structure prediction community makes predictions that are then assessed by independent reviewers. Attendees tested recently developed ab initio protein structure predictions methods during the CASP3 exercises, conducted in December 1998 in Asilomar, California.[29] They presented a number of new techniques that constitute qualitative progress in ab initio prediction with respect to the previous CASPs (held every two years).

Among the best performing ab initio methods was the Rosetta method developed by David Baker and coworkers.[30] This approach works as follows: First, its developers prepared a multiple sequence alignment for the sequence of interest and did secondary prediction. The combined secondary structure predictions and sequence alignments provide the most plausible three- to nine-residue structural fragments extracted (25 fragments for each segment of the query sequence) from the structural database. An algorithm that randomly inserts these three- and nine-residue fragments searches conformational space, and any conformations are scored by a function that contains a hydrophobic burial term, elements of electrostatics, a disulfide bond bias, and a sequence-independent term that evaluates the packing of secondary structure elements. The top 25 (of 1,200 generated) structures frequently contained the proper fold. The best five structures that exhibited a single hydrophobic core are selected by "visual inspection." This could be a drawback because doing a manual evaluation of massive-scale predictions would be difficult. Nevertheless, of 18 targets in CASP3, four predictions proved globally correct (with a root-mean-square deviation range of 4 to 6 Å from native). Furthermore, the majority

of the predictions contained correct fragments.[31]

Other groups also made good predictions for a number of difficult ab initio target proteins at CASP3. Ortiz and colleagues applied a high coordination lattice model that we had first developed, which searched conformational space by an MC-simulated annealing approach.[27,32] The model assumed a 90-basis vector representation of the alpha carbon trace that has a 1.2 Å resolution due to the underlying cubic lattice grid's spacing. Off-lattice single-sphere side chains could assume multiple orientations with respect to the backbone, thereby mimicking the distribution of rotamers for particular amino acids. The model's generic force field consisted of knowledge-based potentials (derived from the statistics of the regularities seen in known protein structures). Additionally, they implemented a weak bias toward predicted secondary structures and weak theoretically predicted long-range contact restraints from correlated mutation analysis in the interaction scheme.[33] They based contact prediction on the analysis of correlated mutations in sequences detected by multiple sequence alignments. For some targets, their approach correctly predicted globally correct fold or large fragments of the structure.

Osguthorpe employed a continuous-space model and sampled conformations with knowledge-based potentials.[34] He correctly predicted substantial fractions of his attempted targets, and his prediction was the best for one of the difficult targets.

Ram Samudrala and coworkers developed a hierarchical procedure that enumerated all compact conformations by using a diamond lattice model that had multiple residues per lattice vertex.[35] They then selected the best structures by fitting the predicted secondary structure fragments to the lattice models. These structures were energy minimized using an all-atom force field and spatial restraints from the lattice models. They scored the optimized structures by a combination of all-atom and residue-based, knowledge-based potentials. They then used distance geometry to generate possible "consensus" models and rebuilt all the atom structures again (optimized and ranked by energy). This method correctly predicted a number of qualitatively correct significant-size protein fragments. This approach's major weakness was perhaps the small fraction of good structures in the initial pool of lattice models.

Harold Scheraga and coworkers developed their force field based on physical principles rather than evolutionary information, which distinguishes their approach from other participants in CASP3.[36] Optimization is performed with conformational space annealing, which narrows the search regions and finds distinct families of low-energy conformations. Then, the lowest energy, reduced model conformations are subsequently converted to the all-atom models and optimized by electrostatically driven MC simulations.[37] For some CASP3 targets, this method produced exceptionally good predictions. The method seemed to perform much better on helical proteins than on β or α/β proteins.

Choice of sampling scheme

In general, the choice of the simulation–optimization algorithm depends on the given study's aim. The study of protein dynamics and folding pathways requires different procedures (and to some extent, different force fields) than those studies designed to identify a protein's native conformation.

MC procedures use a wide spectrum of strategies for conformational updating. In some algorithms, there are global updates of the entire chain; chain growth algorithms are representative of this genre. Other algorithms involve local chain updates involving only a small portion of the chain or a small distance displacement of a larger part of it. Sometimes, the local and global modifications combine in the same algorithm.

The local and global modifications combine in the same algorithm.

If we want to study the kinetics of protein folding, then we need to reproduce the actual process of it. Is there a relationship between the molecular dynamics simulations of a continuous model and a trajectory of an otherwise similar but now discretized (or lattice) model? When a random scheme selects only small, local distance moves, then the dynamics is equivalent to coarse-grained Brownian dynamics, and a given trajectory is the numerical solution of a stochastic equation of motion. Of course, the short-time dynamics on a single elementary move of the discrete model's time scale have no physical meaning. However, the long-time dynamics should be qualitatively correct, albeit with possible distortions of the time scale of various dynamic events. Recent studies show that the MC folding pathways observed in high-coordination lattice models re-

produce the qualitative picture of folding dynamics seen in experiment.[38] Thus, we can use lattice dynamics for meaningful studies of the nature of protein folding pathways and the mechanism of multimeric protein assembly. The validity of studies using discrete models depends more on the protein representation's accuracy and its attendant force field than on a particular sampling scheme. However, some oversimplified discrete models might face serious ergodicity problems—an aspect of simulation that we must carefully examine.

Existing methods do not guarantee that we can find the lowest energy conformation.

We need isothermal simulations at a range of temperatures above, at, and below the folding transition temperature (where 50 percent of the molecules are native and 50 percent are unfolded) to obtain the folding process's thermodynamics. Unfortunately, there is a serious problem associated with the extremely slow relaxation in the low-temperature, dense, globular state where the local barriers are high, thus standard sampling becomes ineffective. This renders straightforward molecular dynamics or canonical MC algorithms prohibitively expensive. We can surmount such problems by using properly designed local moves that can "jump over" these high local energy barriers.

Multicanonical (or Entropy Sampling Monte Carlo—ESMC[20]) sampling can provide more complete data on folding thermodynamics.[25,39,40] Because they use differently defined transition probabilities, energy barriers are substituted by entropic barriers. These simulations offer the advantage of an objective means of establishing when the simulation has converged over a given energy range and from a single series of simulations. It is possible to obtain an estimation of all thermodynamic functions (energy, free energy, and entropy) over a wide range of temperatures. However, the cost of such computations grows rapidly with system size.

Rather than characterizing the full thermodynamics, a simpler task is to find the lowest energy state. This is important because the thermodynamic hypothesis postulates that native proteins are in the conformational energy's global minimum.[41] Researchers have developed a variety of strategies to obtain this global minimum problem including the diffusion equation method, which deforms the energy surface until a single minimum remains. When traced back to the original energy surface, this corresponds to the global energy minimum on the nondeformed surface. For relatively simple but nontrivial systems, this method works well, but for more complex situations, existing methods do not guarantee that we can find the lowest energy conformation.

Simulated annealing, ESMC, minimization, GAs, and the combination of GAs with MC sampling have successfully found the near-native conformations of reduced models of small proteins.[12,13,20,42,43] Recently, many studies have compared the efficiency of various MC strategies for finding a protein model's global minimum.[44] The most straightforward approach is simulated annealing, where the system starts out at a relatively high temperature that gradually lowers until it's below the folding transition temperature. If, on repeated runs (starting from different initial states), we cover the same conformation, we can assume that there is a good chance we have located the global minimum. However, for difficult problems, simulated annealing runs (or at least a substantial fraction of them) become trapped in local energy minima that could be far from the properly folded state. Unfortunately, there is no simple test of convergence in the simulated annealing method. Modifying the transition acceptance criteria could considerably improve the simulated annealing's efficiency. For instance, we could perform local minimization before and after the transition and then apply the Metropolis criterion to the locally lowest energy pairs or conformations. This way, the sampling procedure can avoid visits to a large fraction of irrelevant local states.

In contrast to simulated annealing, sampling techniques that use the multicanonical ensemble have convergence tests. In ESMC,[20] the system entropy estimation is constructed by a sampling process controlled by the density of states of particular discretized energy levels. When converged, all energy levels, including the lowest one, should be sampled with the same frequency. The ESMC method is "quasi-deterministic"—meaning the data from the preceding simulations could help improve successive run accuracy. In principle, ESMC should find the lowest energy state, but in practice, the energy spectrum near the lowest energy state could have large entropy barriers, the lowest energy state might not be detected, and this region might not be converged. We could accelerate the convergence rate by artificially deforming the entropy

curve versus energy in the less important, high-energy range.

The Replica Exchange Monte Carlo (REMC) method[45] has a different philosophy. Here, we simulate many copies with a standard Metropolis scheme at various temperatures spanning from high to low. Occasionally, the replicas are randomly swapped according to a criterion that depends on temperature difference and the energy difference. Thus, the low-energy conformations at a higher temperature could move to a lower temperature. At high temperatures, the energy barriers could be surmounted easily, while at low temperatures, the vicinities of the energy "valleys" are efficiently sampled.

Comparing the computational expense of finding the lowest energy state for a simple protein-like copolymer model shows that REMC is much more efficient than MC-based simulated annealing protocol despite the fact that we must simulate multiple copies of the system. The REMC method also finds the low-energy conformations many times faster than ESMC. Furthermore, due to REMC's efficient sampling, we could use samples at various temperatures for the "umbrella"-type estimation of the density of states as a function of energy from which we can obtain all thermodynamic quantities.

Thus for protein structure prediction, we have a variety of tools for searching the conformational space. The key issue is how can such tools be exploited for successful protein structure prediction.

Prospector

We now turn to the practical question of how one goes about predicting protein structure. Given a protein sequence of unknown structure, most people typically run PSI-BLAST[46] over sequences from structures in the protein data bank.[47–52] Then, if this fails, a threading program in an attempt to identify a significant probe-template match is used. Even if successful, for non-trivial cases, query sequence alignments could be in error. Additionally, there could be gaps in the alignment as well as long unaligned regions. If both methods fail, then ab initio folding is the requisite structure prediction method. Ideally, we want a unified approach that automatically treats these possibilities. Let's look at our recently developed unified approach and then concentrate on the ab initio component.

First, we run our threading algorithm, Prospector,[47] and establish if there is a significant query sequence-template structure match. If so, there are soft biases to the template by a generalized comparative modeling approach that involves ab initio folding in the vicinity of the template in a reduced protein model, the SIde CHain Only Model (SICHO) where each residue is described by a single interaction center located at center of mass of the side chains along with the backbone alpha carbon.[48,49] We use REMC to explore conformational space, but threading can also provide predicted secondary structure and tertiary contacts that are not restricted to the template structure but that we can extract from other structures. This allows for fold prediction in unaligned regions of the query sequence. Conversely, when there is no significant match to a template, the predicted secondary structure and tertiary contacts extracted from threading (onto templates that do not have the query sequence's global structure) are passed to an ab initio folding algorithm that uses the same reduced protein model, but now there are no templates. Then, we cluster the resulting structures,[50] add atomic detail, then use a pairwise atomic potential to better select structures (including low RMSD structures that do not cluster),[51] the structures are again selected, and than present the results.

There could be gaps in the alignment as wellas long unaligned regions.

Summary of CASP4 prediction results

Last year, the next CASP, CASP4 was held. We begin by describing how we did in CASP4. For difficult targets, classified by the CASP4 assessors as "new folds," our method failed to correctly predict the entire structure. Often, Prospector correctly predicted the fold's structure elements as well as supersecondary structure elements, but these elements had topological errors that led to a large overall root-mean-square deviation from the experimental structures. In other cases, we obtained an accurate fold corresponding to the native structure's mirror-image topology.

Sometimes we obtained accurate models in spite of the fact that our threading procedure did not recognize remotely similar folds present in the protein data bank.[52] For example, as shown in Figure 1a for target T0102 (a cyclic 70 amino acid protein), our procedure produced a good model with a coordinate root-mean-square deviation of 3.6 Å from native for the first model (of a maximum of five allowed) submitted. Other

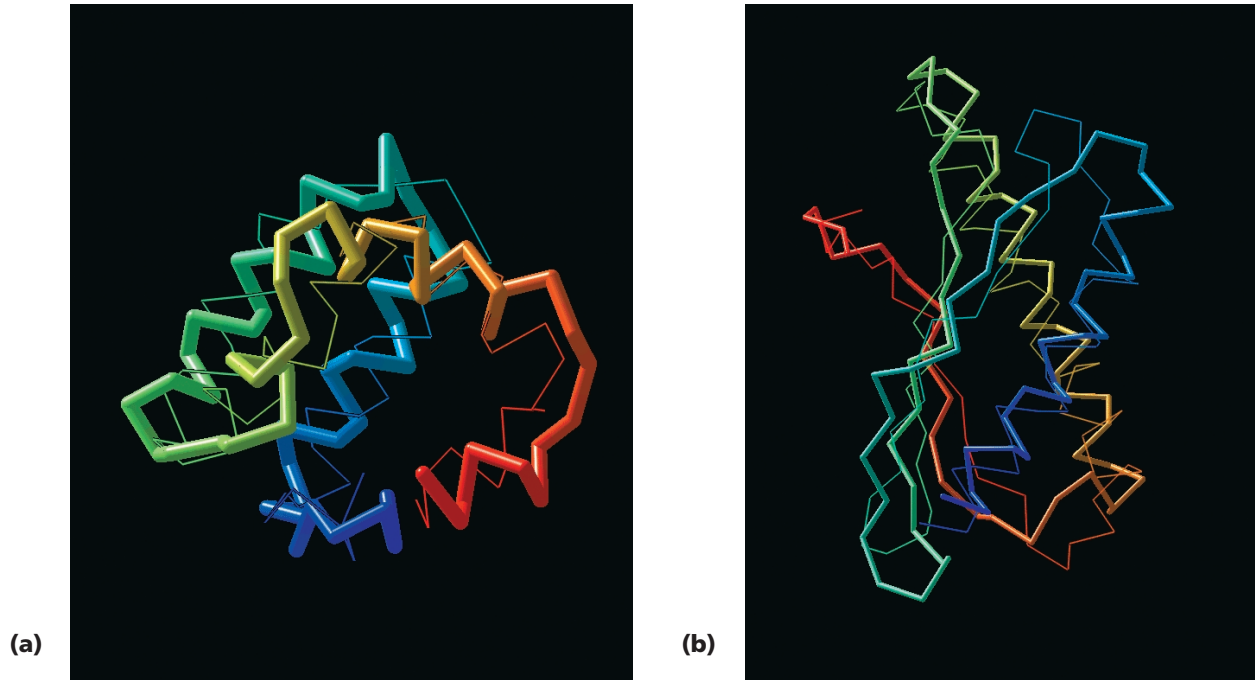

Figure 1. Comparison of the (a) predicted and (b) experimental structures for the CASP4 targets A. T0102 with an RMSD of 3.6 Å from native, and B. T0110, where the predicted structure has an RMSD of 4.2 Å from native.

groups produced models of comparable quality in the range of 4.0 to 4.3 Å from native.

For T0110, shown in Figure 1b, (a 95-residue *a/ß* protein of a complicated fold), our ab initio prediction produced the most accurate model, with a root-mean-square deviation of 4.2 Å from native, which was significantly better than those based on fold recognition or alternative ab initio techniques.

Application to a large benchmark

Subsequent to CASP4, we tuned the SICHO model to improve its performance and improved the contact prediction protocol by using additional protein-specific pair potentials. We also improved the sequence profile method, which defines the score as the difference between the sequence in the structure and the reversed sequence in the structure. The latter modification also makes Prospector more sensitive. We selected sequences of 65 representative small single-domain globular proteins as a test set for ab initio folding.[53] The set contained proteins—α/β, α+β, and β-type folds—and 40 proteins randomly chosen from another work.[54] For 47 out of 65 proteins (72.3 percent), at least one cluster centroid in the top five had a root-mean-square deviation below 6.5Å from native. When we used an atomic potential to select structures, Prospector successfully predicted 50 proteins.[51] If we count the best structure, 58 proteins (89.2 percent) had a structure equal to or below 6.5Å. Unfortunately, the lowest energy structures of only 36 proteins satisfy this criterion, which demonstrates the imperfections in our potentials as well as in the practicality of selecting structures by clustering. Often there are pairs of topological mirror-image structures among the obtained cluster centroids. When one of the centroids has the proper fold, we also obtain (in most cases) the topological mirror-image structure.

Feasibility of structural refinement

Our reduced protein model used to assemble topologies has limited resolution. Typically, well-folded structures have a root-mean-square deviation of 2 to 6.5 Å from native. Can we improve such models using a more detailed protein representation and a more exact force field? It appears that sometimes we can refine the models to a resolution close to that of experimental structures. In previous work with similar low-resolution lattice models, researchers successfully refined several structures of leucine zippers to experimental resolution with a root-mean-square deviation of 0.6 Å from native.[28] We subsequently extended these using ESMC to provide a treatment of the GCN4 leucine zipper

folding thermodynamics as well as the prediction of the native state.[55] Furthermore, the CHARMM force field, when supplemented by a generalized born–surface area treatment, is highly correlated with the lattice-based force field. These studies are extremely encouraging, but it is unclear how soon low-resolution to moderate- or high-resolution structure refinement will become routine.

Although the methodology for protein structure prediction is partially successful, it needs further improvement. Prospector, which forms this approach's core, also needs improvement. For example, it currently uses a very simple sequence profile as a scoring function. Clearly, it needs to exploit more powerful and more sensitive sequence profiles.[56] Prospector also generates high-scoring local sequence fragments that are often quite accurate. This information should be incorporated into subsequent threading iterations and could serve as partial seed structures in ab initio folding, akin to Rosetta.[30]

The scaling of various contributions to the interaction scheme is now to a large extent arbitrary and adjusted essentially by trial and error. To achieve more accurate scaling, we plan to employ an automated procedure targeted to generating as strong a correlation as possible between root-mean-square deviation from native. Perhaps we could achieve a significant improvement in the model by introducing approximate electrostatics into the interaction scheme. This should include more implicit treatment of the solvent other than as an intra main chain hydrogen bonds. The goal here is to reduce the model's reliance on predicted tertiary restraints, which almost always dictate folding method's success.

A variety of sparse but rapidly obtained experimental data could increase the accuracy and extend the range of applicability of our structure prediction method. Our ab initio folding procedure employs predicted secondary structure and predicted contacts between side groups. As demonstrated recently for an older version of the SICHO model, knowledge of secondary structure and as few as $N/7$-$N/5$ side chain contacts (where N is the number of residues in the protein) enable the structure assembly for proteins up to 240 residues.[57] The larger the number of known contacts, the better the accuracy of the predicted structures. We could extract such fragmentary structural data from NMR experiments. Structural restraints could also originate from electron microscopy, fluorescence data, or cross-linking experiments. Sometimes mutation experiments can identify residues that are involved with ligand binding or that are in contact. We could easily incorporate information about the spatial arrangement of these residues into the folding algorithm.

Although techniques for the prediction of low-resolution structures have significantly improved, they still have a way to go before structure prediction becomes routine. Nevertheless, this is a laudable goal as low-resolution structures are of considerable utility both in the identification of biochemical function and in ligand docking.[58] Such efforts will have to be applied on a genomic scale if structure-based approaches to function prediction are to play a role in the post genomic era. A number of such efforts are underway and as structure prediction continues to improve, the applications of protein structure prediction methods to entire genomes will become more prevalent.

Acknowledgment

This research was supported in part by NIH grants Nos. GM37408, GM48835, and RR-12255. We greatly appreciate the contributions of Marcos Betancourt, Hui Lu, Daisuke Kihara, Piotr Rotkiewicz, and Michal Boniecki to some of the research described in this review.

References

1. S.F. Altschul et al., "Basic Local Alignment Search Tool," J. Molecular Biology, vol. 215, 1990, pp. 403–410.
2. S. Henikoff and J.G. Henikoff, "Protein Family Classification Based on Searching a Database of Blocks," *Genomics*, vol. 19, 1994, pp. 97–107.
3. J.S. Fetrow and J. Skolnick, "Method for Prediction of Protein Function from Sequence Using the Sequence-to-Structure-to-Function Paradigm with Application to Glutaredoxins/Thioredoxins and T1 Ribonucleases, *J. Molecular Biology*, vol. 281, 1998, pp. 949–968.
4. J. Skolnick and J. Fetrow, "From Genes to Protein Structure and Function: Novel Applications of Computational Approaches in the Genomic Era," *Tibtech*, vol. 18, 2000, pp. 34–39.
5. J. Bonanno, "Structural Genomics," *Current Biology*, vol. 9, no. 23, 1999, pp. R871–872.

6. R. Sanchez and A. Sali, "Evaluation of Comparative Protein Structure Modeling by MODELLER-3," *Proteins*, vol. 1, 1997, pp. 50–58.
7. D.T. Jones, "GenTHREADER: An Efficient and Reliable Protein Fold Recognition Method for Genomic Sequences," *J. Molecular Biology*, vol. 287, no. 4, 1999, pp. 797–815.
8. M.J. Sternberg et al., "Progress in Protein Structure Prediction: Assessment of CASP3," *Current Opinions in Structural Biology*, vol. 9, no. 3, 1999, pp. 368–373.
9. D. Bashford and D.A. Case, "Generalized Born Models of Macromolecular Solvation Effects," *Ann. Rev. Physical Chemistry*, vol. 51, 2000, pp. 129–152.
10. M. Levitt and A. Warshel, "Computer Simulation of Protein Folding," *Nature*, vol. 253, 27 Feb. 1975, pp. 694–698.
11. C. Wilson and S. Doniach, "A Computer Model to Dynamically Simulated Protein Folding: Studies with Crambin," *Proteins*, vol. 6, 1989, pp. 193–209.
12. S. Sun, "Reduced Representation Model of Protein Structure Prediction: Statistical Potential and Genetic Algorithms," *Protein Science*, vol. 2, 1993, pp. 762–785.
13. J.T. Pedersen and J. Moult, "Ab Initio Protein Folding Simulations with Genetic Algorithms: Simulations on the Complete Sequence of Small Proteins," *Proteins*, vol. 1, 1997, pp. 179–184.
14. N. Go and H. Taketomi, "Respective Roles of Short- and Long-Range Interactions in Protein Folding," *Proc. Nat'l Academy of Science*, 1978, pp. 559–563.
15. W.R. Krigbaum and S.F. Lin, "Monte Carlo Simulation of Protein Folding Using a Lattice Model," *Macromolecules*, vol. 15, 1982, pp. 1135–1145.
16. A. Kolinski and J. Skolnick, "Monte Carlo Simulations on an Equilibrium Globular Protein Folding Model," *Proc. Nat'l Academy of Science*, 1986, pp. 7267–7271.
17. J. Skolnick and A. Kolinski, "Computer Simulations of Globular Protein Folding and Tertiary Structure," *Ann. Rev. Physical Chemistry*, vol. 40, 1989, pp. 207–235.
18. J. Skolnick and A. Kolinski, "Simulations of the Folding of a Globular Protein," *Science*, vol. 250, 1990, pp. 1121–1125.
19. H.S. Chan and K.A. Dill, "Polymer Principles in Protein Structure and Stability," *Ann. Rev. Biophysics and Biophysical Chemistry*, vol. 20, 1991, pp. 447–490.
20. M.-H. Hao and H.A. Scheraga, "Monte Carlo Simulations of a First-Order Transition for Protein Folding," *J. Physical Chemistry*, vol. 98, 1994, pp. 4940–4948.
21. L. Mirny and E. Shakhnovich, "Protein Folding Theory: From Lattice to All-Atom Models," *Ann. Rev. Biophysics and Biomolecular Structures*, vol. 30, 2001, pp. 361–396.
22. V.G. Dashevskii, "Lattice Model of Three-Dimensional Structure of Globular Proteins," *Molekulyarnaya Biologiya*, vol. 14, no. 1, 1980, pp. 105–117.
23. D.G. Covell, "Folding Protein α-Carbon Chains into Compact Forms by Monte Carlo Methods," *Proteins*, vol. 14, 1992, pp. 409–420.
24. D.A. Hinds and M. Levitt, "A Lattice Model for Protein Structure Prediction at Low Resolution," *Proc. Nat'l Academy of Science*, 1992, pp. 2536–2540.
25. A. Kolinski and J. Skolnick, *Lattice Models of Protein Folding, Dynamics and Thermodynamics*, //**au: publisher?**//, 1996.
26. J. Skolnick et al., "A Method for Prediction of Protein Structure from Sequence," *Current Biology*, vol. 3, 1993, pp. 414–423.
27. A. Kolinski and J. Skolnick, "Monte Carlo Simulations of Protein Folding, II: Application to Protein A, ROP, and Crambin," *Proteins*, vol. 18, 1994, pp. 353–366.
28. M. Vieth et al., "Prediction of the Folding Pathways and Structure of the GCN4 Leucine Zipper," *J. Molecular Biology*, 1994, pp. 361–367.
29. J. Moult et al., "Critical Assessment of Methods of Protein Structure Prediction (CASP): Round III," *Proteins*, vol. 3, 1999, pp. 2–6.
30. K.T. Simons et al., "Ab Initio Protein Structure Prediction of CASP III Targets Using ROSETTA," *Proteins*, vol. 3, 1999, pp. 171–176.
31. D.T. Jones, "Successful Ab Initio Prediction of the Tertiary Structure of NK-Lysin Using Multiple Sequences and Recognized Supersecondary Structural Motifs," *Proteins*, vol. 1, 1997, pp. 185–191.
32. A.R. Ortiz et al., "Ab Initio Folding of Proteins Using Restraints Derived from Evolutionary Information," *Proteins*, vol. 3, 1999, pp. 177–185.
33. A.R. Ortiz, A. Kolinski, and J. Skolnick, "Nativelike Topology Assembly of Small Proteins Using Predicted Restraints in Monte Carlo Folding Simulations," *Proc. Nat'l Academy Science*, 1998, pp. 1020–1025.
34. D.J. Osguthorpe, "Improved Ab Initio Predictions with Simplified Flexible Geometry Model," *Proteins*, vol. 3, 1999, pp. 186–193.
35. R. Samudrala et al., "Ab Initio Proteins Structure Prediction Using a Combined Hierarchical Approach," *Proteins*, vol. 3, 1999, pp. 194–198.
36. J. Lee et al., "Calculation of Protein Conformation by Global Optimization of a Potential Energy Function," *Proteins*, vol. 3, 1999, pp. 204–208.
37. D.R. Ripoll, A. Liwo, and H.A. Scheraga, "New Developments of the Electrostatically Driven Monte Carlo Method: Test on the Membrane-Bound Portion of Melittin," *Biopolymers*, vol. 46, 1988, pp. 117–126.
38. A. Rey and J. Skolnick, "Comparison of Lattice Monte Carlo Dynamics and Brownian Dynamics Folding Pathways of α-Helical Hairpins," *Chemical Physics*, vol. 158, 1991, pp. 199–219.
39. U.H.E. Hansmann and Y. Okamoto, "Prediction of Peptide Conformation by Multicanonical Algorithm: New Approach to the Multiple Minima Problem," *J. Computational Chemistry*, vol. 14, 1993, pp. 1333–1338.
40. A. Kolinski, W. Galazka, and J. Skolnick, "Monte Carlo Studies of the Thermodynamics and Kinetics of Reduced Protein Models: Application to Small Helical, b and a/b Proteins," *J. Chemical Physics*, vol. 108, 1998, pp. 2608–2617.
41. C.B. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science*, vol. 181, 1973, pp. 223–230.
42. H.A. Scheraga and M.-H. Hao, "Entropy Sampling Monte Carlo for Polypeptides and Proteins," *Advanced Chemical Physics*, vol. 105, 1999, pp. 243–272.
43. T. Dandekar and P. Argos, "Identifying the Tertiary Fold of Small Proteins with Different Topologies form Sequence and Secondary Structure Using the Genetic Algorithm and Extended Criteria Specific for Strand Regions," *J. Molecular Biology*, vol. 256, 1996, pp. 645–660.
44. U.H.E. Hansmann and Y. Okamoto, "Numerical Comparison of Three Recently Proposed Algorithms in the Protein Folding Problem," *J. Computational Chemistry*, vol. 18, 1997, pp. 920–933.
45. R.H. Swendsen and J.S. Wang, "Replica Monte Carlo Simulation of Spin Glasses," *Physical Rev. Letters*, vol. 57, no. 21, 1986, pp. 2607–2609.
46. S.F. Altschul and E.V. Koonin, "Iterated Profile Searches with PSI-BLAST: A Tool for Discovery in Protein Databases," *Trends in Biochemical Science*, vol. 23, no. 11, 1998, pp. 444–447.
47. J. Skolnick and D. Kihara, "Defrosting the Frozen Approximation: A New Approach to Threading, *Proteins*, vol. 42, 2001, pp. 319–331.
48. J. Skolnick et al., "Ab Initio Protein Structure Prediction via a Combination of Threading, Lattice Folding, Clustering, and Structure Refinement," to appear in *Proteins*, 2001.
49. A. Kolinski et al., "Generalized Comparative Modeling (GENECOMP): A Combination of Sequence Comparison, Threading, and Lattice Modeling for Protein Structure Prediction and Refinement," to appear in *Proteins*, 2001.

50. M.R. Betancourt and J. Skolnick, "Finding the Needle in a Haystack: Educing Native Folds from Ambiguous Ab Initio Protein Structure Predictions," *J. Computational Chemistry*, vol. 22, 2001, pp. 339–353.
51. H. Lu and J. Skolnick, "A More Specific Distant Dependent Atomic Knowledge Based Potential for Protein Structure Prediction," to appear in *Proteins*, 2001.
52. H.M. Berman, "The Past and Future of Structure Databases," *Current Opinions in Biotechnology*, vol. 10, no. 1, 1999, pp. 76–80.
53. D. Kihara et al., "TOUCHSTONE: An Ab Initio Protein Structure Prediction Method that Uses Threading-Based Tertiary Restraints," to appear in *Proc. Nat'l Academy of Science*, 2001.
54. K.T. Simons, C. Strauss, and D. Baker, "Prospects for Ab Initio Protein Structural Genomics," *J. Molecular Biology*, vol. 306, no. 5, 2001, pp. 1191–1199.
55. D. Mohanty, A. Kolinski, and J. Skolnick, "De Novo Simulations of the Folding Thermodynamics of the GCN4 Leucine Zipper," *Biophysical J.*, vol. 77, no. 1, 1999, pp. 54–69.
56. L. Rychlewski et al., "Comparison of Sequence Profiles: Strategies for Structural Predictions Using Sequence Information," *Protein Science*, vol. 9, no. 2, 2000, pp. 232–241.
57. A. Kolinski, "Assembly of Protein Structure from Sparse Experimental Data: An Efficient Monte Carlo Model," *Proteins*, vol. 32, 1998, pp. 475–494.
58. M. Wojciechowski and J. Skolnick, "Docking of Small Ligands to Low-Resolution and Theoretically Predicted Receptor Structures," to appear in *J. Computational Chemistry*, 2001.

Jeffrey Skolnick is the Director of Computational and Structural Biology at the Donald Danforth Plant Science Center. His research interests are in computational biology, protein structure and function prediction, lattice-based approaches to protein tertiary structure prediction, the simulation of membranes and membrane peptides, and bioinformatics. He received his PhD in polymer statistical mechanics from Yale University. Contact him at the Laboratory of Computational Genomics, Danforth Plant Science Center, 893 N. Warson Rd., Creve Coeur, MO 63141; skolnick@danforthcenter.org.

Andrzej Kolinski is a Member at the Donald Danforth Plant Science Center and Head of the Theory of Biopolymers Laboratory at the University of Warsaw, Poland. He has received the International Scholar's Award of the Howard Hughes Medical Institute and is a three-time recipient of the Prize of Polish Ministry of Higher Education. He received a PhD in chemistry from the University of Warsaw. Contact him at the Laboratory of Computational Genomics, Danforth Plant Science Center, 893 N. Warson Rd., Creve Coeur, MO 63141; kolinski@danforthcenter.org.